Email Analytics System

Email Analytics System allows us to understand the email traffic by date, hour and year as well as graph analysis.

In this project, we'll learn how we can use Apache Sqoop, Apache HDFS, Apache Oozie, and Apache PIG to design an end-to-end data pipeline that will enable us to analyze Email data. Following tasks need to be accomplished in order:

MileStone1: Download the Enron mysql dump(email data) from following link and import the data into mysql database.

MileStone2: Analyze the dataset in mysql tables to get overview of imported data

MileStone3: Import the data from related mysql tables into HDFS using Sqoop Tool.

MileStone4: Perform the following Transformation steps on HDFS data:

a) Load the Message and RecipientInfo data

b) Join the above datasets to get unified Email data with following fields:

message_id, date, from, to, ccs, bccs, subject, body

c) The date must be transformed from database custom format to ISO format.

d) Transform the data into AVRO format

MileStone5: Develop PIG/HIVE/MR jobs to develop following basic statistics from AVRO enron data:

- **Mails per day of the week**: From Monday-Sunday, mails on each day
- **Mails per hour of the day**: From 0 hours to 23 hours, mails by hour
- **Mails per month**: Emails received by month across all years
- **Mails per year:** Emails received across all year
- **Mails received**: Top 50 people who received more email
- **Mails sent**: Top 50 people who sent more email

MileStone6: Create the target tables in mysql for storing above statistics and export the results from HDFS to mysql tables. UI tools will read from mysql tables and provides visualization.